A Classroom Experiment on Effort Allocation under Relative Grading

Andy Brownback

August 14, 2014

Abstract

Grading on the curve, or relative grading is one of the most commonly used grading schemes in education. Under this scheme, each student’s grade is determined as a function of his or her percentile rank in the class. According to the law of large numbers, this mechanism is sensitive to changes in the size of the class. As a class size grows, the percentile ranks of students in a class draw closer to their respective percentile ranks in the student population, changing the incentives for students. I develop a model of this environment that incorporates strategic effort in order to predict behavior for students with different abilities. I test this model in a large-scale classroom experiment that measures student effort on 10 online quizzes. I randomly assign students to a “grading cohort” of either 10 or 100 students, where students receive high grades if their performance is in the top 70 percent of their cohort, and test the connection between the size of the grading cohort and the effort students exert. Larger grading cohorts have less variance in the draw of students, and elicit greater average effort. Smaller grading cohorts have greater variance in the draw of students, and elicit more effort from the lowest ability students. Students in the bottom 30 percent exert greater effort on average in the larger grading cohort, an allocation failure consistent with several possible behavioral biases. I discuss the welfare implications of the observed patterns of effort and make policy prescriptions for different objectives of a classroom designer.
1 Introduction

Under a relative or “curved” grading scheme, a student’s academic performance is evaluated relative to her peers, and her grade is determined based on her percentile rank within her comparison group. Importantly, these grades are determined independently from any absolute measures of performance. The advantage of not needing to determine a cardinal measure of academic performance has helped establish these grading mechanisms as a fixture in many university classrooms and law schools.\textsuperscript{1,2} Indeed, mechanism designers across many areas of education have employed relative awarding schemes in competitions for scholarships, college admissions, and even teacher pay.\textsuperscript{3,4,5}

In its simplest form, relative grading assigns grades to students based only on their percentile rank within the class. For example, a professor may only award “A” grades to students who are at or above the 80 percentile of performers. Call the quantiles that serve to distinguish different grade levels “critical cutoffs.”

Consider an example, Texas HB 588, which grants automatic admission to any Texas state university—including the University of Texas at Austin—to all Texas high school seniors who graduate in the top 10 percent of their high school class.\textsuperscript{6} The incentive structure of this policy mirrors that of a curved grading scheme with a single critical cutoff at the 10th percentile. Like many other curved grading schemes, this policy is invariant to the number of students in a given graduating class despite the fact that graduating classes can vary by orders of magnitude.\textsuperscript{7}

\textsuperscript{1}Though frequently discussed, good statistics on the percentage of undergraduate courses graded on the curve are hard to come by.

\textsuperscript{2}Mroch (2005) estimates that 79% of law schools standardize scores according to a grading curve.

\textsuperscript{3}Missouri’s Bright Flight scholarship program awards scholarships to the top 3% of high school seniors based on ACT or SAT scores.

\textsuperscript{4}In California, the top 9% of graduating seniors are guaranteed admission to one of the University of California campuses. In Kansas, the top 33% are guaranteed access to the state college of their choice. In Texas, the top 10% are offered similar incentives.

\textsuperscript{5}North Carolina Senate Bill 402 section 9.6(g) grants favorable contracts to the top 25% of teachers at each school as evaluated by the school’s administration.

\textsuperscript{6}The bill was modified in 2009 to stipulate that the University of Texas at Austin may cap the number of students admitted under this measure to 75% of in-state freshman students.

\textsuperscript{7}Plano East High School in Plano, TX has an enrollment of 6,015 students, while Valentine High School
Under curved grading, a student’s outcomes depend on her comparison group. Call the draw of students in a comparison group a “grading cohort.” The law of large numbers says that the larger is a draw of students from a distribution, the more that draw resembles the distribution itself. Since the incentives are determined by the distribution of students, the size of the grading cohort must affect the incentives. To put the incentives in context, the larger is the Texas high school graduating class, the closer is the distribution of students in it to the population distribution of Texas high school seniors. Small schools are more likely to draw a senior class of outlying students than large schools, resulting in grading cohorts with higher variance. A similar intuition holds for university courses, grant applications, and tenure-track positions, to list a few examples.

This paper presents a theoretical and experimental treatment of a simplified grading curve where students either receive a high or low grade. The theoretical model helps formalize the ways in which the incentive structure of relative grading schemes like HB 588 are sensitive to changes in the number of competitors. The existence of different incentives, however, does not ensure that participants’ behavior will respond predictably. Prior research has demonstrated that economic agents often ignore sample size when drawing inference (Tversky and Kahneman, 1971; Kahneman and Tversky, 1973; Rabin, 2002; Rabin and Vayanos, 2010). This paper seeks to address this question outside of the laboratory environment by testing the sensitivity of students’ choices to incentives that change with sample size.

The model in this paper employs rational-agent economic theory to develop equilibria for two identical, grading curves with differently sized grading cohorts. While assigning expected grade maximizing preferences to students may seem like a dubious assumption, this model has the advantage of providing simple, closed form predictions for how changes in a student’s environment can affect his incentives for effort. Where the model fits the data, we will have strong evidence that students can draw inference from the Law of Large Numbers. Where behavior systematically deviates from the theory, alternative theories will need to be tested.

in Valentine, TX has an enrollment of 9 students.
In the discussion section I provide 3 possible behavioral biases that may be responsible for observed deviations from the model.

I conduct an experiment on relative grading in a large class of undergraduates at a major public university. I present students with a pair of quizzes graded according to identical curves, that is, with identical quantiles set as critical cutoffs. For each student, I randomly determine which quiz will be graded based on the 10-Student cohort and which will be graded based on the 100-Student cohort. I observe the amount of time each student spends taking each quiz and use this as a measure of effort. My design allows me to control for the amount of time studied in preparation for the quiz, so I can use the time spent on a given quiz to directly analyze the relationship between the effort exerted on quizzes and the size of the grading cohort. Similar to existing models, my model predicts that effort will increase with the size of the grading cohort. My results confirm this prediction, showing that larger grading cohorts elicit a statistically significantly 30 second increase in effort.

I enrich the analysis by including GPA to control for student ability. The model makes clear predictions about the heterogeneity of incentives across students of different ability levels. First, the model predicts that students whose ability level puts them below the critical cutoff will exert more effort in the smaller grading cohorts, while students above the critical cutoff will do the opposite. Second, the model predicts that the maximum difference in effort between cohort sizes should occur near the critical cutoff on either side.

My results deviate from these predictions in important ways. First, while the lowest ability students exert significantly more effort on quizzes with the smaller grading cohort, students below the critical cutoff exert more effort in the larger grading cohorts on average. Second, the maximum difference in favor of the large grading cohort actually occurs before the critical cutoff, not after it. For students just before the cutoff, this effort allocation fails to take advantage of the greater variance of the smaller grading cohort and may arise from one of several different behavioral biases. If students fail to update about their own ability

---

8 Most notably Moldovanu and Sela 2006
or their grading environments, if they possess inherent overconfidence, or if their preferences are myopically reference dependent, then they may exhibit these allocation failures.

The tension between aggregate effort and effort exerted by specific types of students is central to the debate over optimal classroom design. I propose ways in which the effect of contest size can be addressed within the curved grading mechanism, depending on the designer’s objective function. I also propose different policies targeted at the behavioral biases potentially causing students to misallocate effort across quizzes.

To fix ideas, I refer to “relative” or “curved” grading schemes throughout this paper, but this should not distract from the generality of the results. While my discussion and experiment all exist within the classroom setting, relative awarding schemes are found throughout the modern economy, in job promotion contests, performance bonuses, and lobbying contest, to name a few. I believe that my results generalize to any of these settings for several reasons. To begin, the costs of effort in my experiment, the means of exerting it, and the norms surrounding its exertion are all similar to a professional setting. Additionally, the means of evaluating performance in my experiment closely reflects those of many professional settings. Finally, the heterogeneity of ability in a classroom can be thought to follow very similar patterns as the heterogeneity of professional competencies, since intelligence, motivation, and cognitive ability drive both.

The outline of this paper is as follows. In the following section, I discuss the existing literature on contests in experimental and non-experimental settings, as well as the education literature on performance evaluations. In Section 3 I outline a simple model of the type of relative grading that my experimental subjects face. This model develops the predictions that I will test in the experiment. Section 4 solves the model under a specific set of production and cost functions. The experiment itself is formally presented in Section 5. Section 6 presents the results of the experiment and tests the predictions of the model. Section 7 discusses the

\footnote{For example, in his book \textit{Straight from the Gut}, former GE CEO Jack Welch recommends that managers rank their employees according to a 20-70-10 model of employee vitality. He claims roughly 20\% of employees should be labeled "A" players, 70\% "B" players, and 10\% "C" players. "A" players should be rewarded, and "C" players should be eliminated.}
implications of my results on the optimal design of classroom grading mechanisms as well as other relative awarding schemes. Section 8 concludes the paper.

2 Literature

In addressing the strategic incentives of classroom grading structures, this paper spans three distinct literatures. Its first contribution is within the literature of experimental economics. Accordingly it owes a debt to many prior experimental tests of contests and auctions. Second, the theoretical model has its roots in a long line of theoretical literature on contests. Notably, Becker and Rosen (1992), who modify the tournament structure of Lazear and Rosen (1981) in order to generate predictions for student effort under relative or absolute grading schemes. Third, in modeling and collecting data from a classroom setting, it continues a literature on classroom performance that has been explored in the literatures of development economics and economics of education. I will address each of these three branches of the literature in turn.

2.1 Experimental Economics

Experiments testing effort exertion in different laboratory settings date back to Bull, Schotter and Weigelt (1987), who tested bidding in laboratory rank-order tournaments. Bidders in their experiment approached equilibrium after several rounds of learning. Equilibrium in all-pay auctions in the laboratory has been more elusive. Indeed the vast majority of studies demonstrate overbidding (Potters, de Vries, and van Winden, 1998; Davis and Reilly, 1998; Gneezy and Smorodinsky, 2006; Barut, Kovenock, and Noussair, 2002). Müeller and Schotter (2010) and Noussair and Silver (2006) confirm the overbidding result, but also demonstrate heterogeneous effects across different types of bidders.

The independent effect of contest size on effort has rarely been addressed in laboratory experiments. The majority of studies on the effects of contest size on effort restrict their
focus to either small changes in the size of contest (Harbring and Irlenbusch, 2005) or changes in the size that affected the proportion of winners (Gneezy and Smorodinsky, 2006; Müeller and Schotter, 2010; Barut et al., 2002, List et al., 2014). Andreoni and Brownback (2014) begin to address this gap by providing a framework for evaluating the effects of contest size on bids in a laboratory all-pay auction. Larger contests in this setting are found to generate greater aggregate bidding, greater bidding by high types, and lower bidding by low types. These results reveal the fundamental tension between maximizing aggregate bidding and maximizing bidding by disadvantaged types.

This paper takes the framework of Andreoni and Brownback (2014) out of the laboratory and into a field setting and is, to my knowledge, the only study that directly measures effort as a function of the classroom size and the grading environment. Experiments on classroom size have been conducted to test other outcomes. For example, the state of Tennessee experimented with classroom sizes for kindergarten students (Mosteller, 1995), but the setting was clearly non-strategic. The experimenters in Tennessee were interested in the effect of classroom size on teacher effectiveness and the retention of information by students. In my experiment, I explicitly control for both of these factors in order to exclusively uncover the effect of changing the strategic incentives for output, not the quality of input.

Other economic experiments in education have focused on randomized changes in the inputs to students’ production functions in the form of textbooks (Glewwe, Kremer, and Moulin, 2002), teachers (Banerjee, Jacob, and Kremer, 2004), teacher incentives (Glewwe, Ilias, and Kremer, 2010), remedial education (Banerjee et al. 2014) or additional study material (Glewwe et al., 2004), to name a few. Classroom policy has been the subject of experimentation as well. Studies have explored the responsiveness of effort to mandatory attendance policies (Chen and Lin, 2008; Dobkin, Gil, and Marion, 2010) or different grading policies (Czibor et al., 2014).
2.2 Microeconomic Theory

The contest theory literature began in order to capture the incentives for rent-seeking (Tullock, 1967 and Krueger, 1974), but has evolved into a more general branch of research that considers various environments with costly effort and uncertain payoffs. The three models most frequently used are the all-pay auction (Hirshleifer and Riley, 1978; Hillman and Samet, 1987; Hillman and Riley, 1989), the Tullock contest (Tullock, 1980), and the rank-order tournament (Lazear and Rosen, 1981).

Hillman and Riley (1989) and Baye, Kovenock, and de Vries (1993) use the all-pay auction model to explore the incentives for rent-seeking in politics. Amann and Leininger (1996) introduce incomplete information about the types of opponents a given player faces, generating a pure-strategy equilibrium effort function. Baye, Kovenock, and de Vries (1996) fully characterize the equilibrium of the all-pay auction and demonstrate that a continuum of equilibria exist. Moldovanu and Sela (2001) develop a model of optimal contest architecture for designers with different objective functions. For a comprehensive theoretical characterization of all-pay contests that incorporates many of the existing models into one framework, see Siegel (2009). For an exhaustive survey of the experimental literature on contests and auctions, refer to Dechenaux, Kovenock, and Sheremeta (2012).

This paper is motivated by the way in which the size of a contest may chance its incentives. This intuition is captured in a more general theoretical framework by Moldovanu and Sela (2006), who demonstrate the single-crossing property of the symmetric equilibrium effort functions of differently sized contests. Olszewski and Siegel (2013) provide similar results about equilibria in a general class of contests with a large but finite number of participants.

2.3 Education Literature

In the economics of education literature, grading mechanisms are often designed without consideration of any strategic interaction between students. For example, Costrell (1994) explores the endogenous selection of grading standards by policy makers who wish to maxi-
mize social welfare, subject to students who best respond to those standards. Betts (1998) expands this framework to include heterogeneous students. Betts and Grogger (2003) then look at the impact of grading standards on the distribution of students.

Both Paredes (2012) and Dubey and Geanakoplos (2010) compare incentive across different methods of awarding grades. Paredes (2012) considers a switch from an absolute to a relative grading scheme, while Dubey and Geanakoplos (2010) consider the optimal coarseness of the grades reported when students gain utility from their relative rank in the class.

Kokkelenberg, Dillon, and Christy (2006) look at the effect of class size on the grades awarded to individual students in a non-strategic environment, and find a strong negative relationship. The independent effect of class size on strategically interacting students, however, remains unstudied. Contest size is often taken as given or assumed to be determined exogenously. In this paper, I demonstrate that classroom size plays an integral role in a student’s selection of effort when grading on the curve. I additionally show that any classroom designer hoping to optimize student outcomes will need to take into consideration its heterogeneous effects on students of different abilities.

3 Model

In this section, I develop a model of competitive interaction in a classroom graded on the curve. To do this, I make certain assumptions about the preferences and production functions of the students. I will pursue generality as much as possible while maintaining a tractable model that can demonstrate the intuition for the theoretical effect that the law of large numbers has on the effort choices of students as the enrollment of a class grows. In my analysis I will borrow heavily from the independent private value auction literature (Vickrey, 1961), and the all-pay auction literature (Baye, Kovenock, and de Vries, 1993).

Suppose there are $N$ students exerting costly effort in order to increase their chances of
winning one of $M \equiv P \times N$ prizes in the form of high grades. These high grades are awarded on a strictly relative basis to the highest $M$ performers in this $N$ student grading cohort.

Suppose each student has an ability, $a_i$, that is independently drawn from the cumulative distribution function, $F(a_i)$. Students are evaluated at each period, $t$, based on their academic output, or “score,” $s_{i,t}$ from that period. Output is determined as a function of effort, $e_{i,t}$, and ability. Suppose this production function is defined as

$$s_{i,t} = S(e_{i,t}; a_i).$$

(1)

Assume that academic output is strictly increasing in all inputs. That is, $\frac{\partial S}{\partial E} > 0$ and $\frac{\partial S}{\partial a} > 0$.

### 3.1 Student’s Utility

The expected utility of a student is be determined by the likelihood of receiving a high grade given her choice of academic output minus the cost of output. Normalize the value of receiving a high grade to one. Heterogeneity across students with different ability levels is now captured by the effort required to generate a given level of output. Thus, represent a student’s utility as

$$U(e_{i,t}; a_i) = \Pr \left( S(e_{i,t}; a_i) \geq \bar{S} \right) - C(e_{i,t}),$$

(2)

where $\bar{S}$ represents the minimum output required to receive a high grade and $C(e_{i,t})$ represents the cost of a student’s effort choice. Assume that the cost of effort is identical across students.$^{10}$ This implies that a student’s cost of a given level of output decreases in her ability.

The optimal score is a function of $P$ and $N$. Therefore, represent the cumulative distribution of scores chosen by students in the population as $G(s_{i,t}, N, P)$. Since the minimum

---

$^{10}$This assumption will not affect the predictions for my experiment, since I control for student-specific costs of effort.
score required for a high grade is dependent on the distribution of scores chosen by other students in the grading cohort, we can characterize it with the following order statistic,

\[
\Pr(s_{i,t} \geq S) = \sum_{j=(1-P) \times N}^{N-1} \left( \frac{(N-1)!}{(N-1-j)! \times j!} \right) G(s_{i,t}; N, P)^j \\
\times (1 - G(s_{i,t}; N, P))^{N-1-j}.
\]  

(3)

Substituting Equation (3) into Equation (2) provides the complete specification of the utility function

\[
U(e_{i,t}; a_i) = \sum_{j=(1-P) \times N}^{N-1} \left( \frac{(N-1)!}{(N-1-j)! \times j!} \right) G(S(e_{i,t}, a_i); N, P)^j \\
\times (1 - G(S(e_{i,t}, a_i); N, P))^{N-1-j} - C(e_{i,t}).
\]  

(4)

In Section 4, I use a simple Cobb-Douglas production function for academic output to complete the specification for Equation (4). I then solve for the student’s effort as a function of her ability in a symmetric Bayesian Nash Equilibrium. I leave the general treatment of equilibrium effort for the appendix.

4 A Simple Example: Cobb-Douglas Production Function

In this section, I will solve the model of Section 3 for a specific set of production and cost functions. These simplifications provide tractability, maintain the key insights of the model, and illustrate generic qualities of equilibria in this setting. In particular, this example captures the heterogeneity in the response to grading cohort size.
Suppose that a student’s academic output is characterized by the Cobb-Douglas production function mapping ability and effort into a score

\[ s_{i,t} = e_{i,t} \times a_i , \]  

with constant, homogeneous marginal costs of effort

\[ C(e_{i,t}; a_i) = c \times e_{i,t} . \]  

This production function is appealing for its simplicity and ability to capture an intuitive complementarity between ability and effort. Constant, homogeneous marginal costs are appealing in that they allow the costs of a given score to be weakly decreasing in ability.

Now, rewrite Equation (4) as

\[ U(e_{i,t}; a_i) = \sum_{j=(1-P)N}^{N-1} \left( \frac{(N-1)!}{(N-1-j)! \times j!} \right) G(e_{i,t}a_i; N, P)^j \times (1 - G(e_{i,t}a_i; N, P))^{N-1-j} - c \times e_{i,t} . \]  

Consider only symmetric equilibria. In the appendix, I demonstrate that the optimal score is monotonic in ability in any symmetric equilibrium. In addition to monotonicity, it is straightforward to show that scores must also be continuous in ability.\(^{11}\) All continuous, monotonic functions are invertible, meaning that there exists a function that maps a given student’s score back onto the student’s ability implied by that score. Given that the equilibrium scores depend on \(N\) and \(P\), this inverse function, too, depends on these parameters. This defines the function \(A(s_{i,t}; N, P) \equiv S^{-1}(s_{i,t}; N, P).\)

Under monotonicity, Equation (3), can be expressed as an order statistic involving the

\(^{11}\)Suppose not. With discontinuities within the support \(G(e_{i,t})\), some students would be failing to best respond. A student scoring just above the discontinuity would be able to increase his expected utility by lowering his effort, which would lower his score and his costs, up until the discontinuity has vanished.
CDF of ability, $F(a_i)$, rather than the CDF of effort. The relevant order statistic now gives
the probability that the ability level implied by a student’s score is higher than the ability
levels implied by the scores of $N(1-P)$ other students in her grading cohort,

$$
\Pr(s_{i,t} \geq \bar{S}) = \sum_{j=(1-P)\times N}^{N-1} \left[ \frac{(N-1)!}{(N-1-j)! \times j!} \right] F(A(s_{i,t}; N, P))^j 
\times (1 - F(A(s_{i,t}; N, P)))^{N-1-j}.
$$

(8)

Substituting in the production function for academic output completes the student’s
utility function

$$
U(e_{i,t}; a_i) = \sum_{j=(1-P)\times N}^{N-1} \left[ \frac{(N-1)!}{(N-1-j)! \times j!} \right] F(A(e_{i,t}a_i; N, P))^j 
\times (1 - F(A(e_{i,t}a_i; N, P)))^{N-1-j} - ce_{i,t}.
$$

(9)

Maximizing Equation (9) with respect to the choice variable, $e_{i,t}$, yields the first-order
condition

$$
\frac{\partial U(e_{i,t}; a_i)}{\partial e_{i,t}} = \sum_{j=(1-P)\times N}^{N-1} \left[ \frac{(N-1)!}{(N-1-j)! \times j!} \right] \left\{ j \times F(A(e_{i,t}a_i; N, P))^{j-1} 
\times f(A(e_{i,t}a_i; N, P)) \times A'(e_{i,t}a_i) \times a_i \right\} 
\times (1 - F(A(e_{i,t}a_i; N, P)))^{N-1-j} 
- \left[ F(A(e_{i,t}a_i; N, P))^j \times (N-1-j) \right] \times (1 - F(A(e_{i,t}a_i; N, P)))^{N-2-j} 
\times f(A(e_{i,t}a_i; N, P)) \times A'(e_{i,t}a_i) \times \alpha a_i \} \equiv c.
$$

13
At equilibrium, the ability implied by a student’s score must equal that student’s ability. That is, $A(s_{i,t}) = a_i$. Substituting this into the first order condition and solving for $\frac{1}{A(s_{i,t})}$ results in

$$
\frac{a_i}{c} f(a_i; N, P) \sum_{j=(1-P)\times N}^{N-1} \left( \frac{(N-1)!}{(N-1-j)! \times j!} \right) \left\{ j \times F(a_i; N, P) \right\}^{j-1} (1 - F(a_i; N, P))^{N-1-j} - \left[ (N-1-j)F(a_i; N, P) \right]^{j} (1 - F(a_i; N, P))^{N-2-j} \right\} \right) \equiv \frac{1}{A'(s_{i,t}; N, P)}. \tag{10}
$$

Since $A(s_{i,t}; N, P) \equiv S^{-1}(s_{i,t}; N, P)$, it holds that $\frac{1}{A'(s_{i,t}; N, P)} = S'(S^{-1}(s_{i,t}))$, implying that $\frac{1}{A'(s_{i,t}; N, P)} = S'(a_i)$. Substituting this equality into Equation (10) yields the following first order differential equation defining the equilibrium score as a function of ability,

$$
S'(a_i; N, P) \equiv \frac{a_i}{c} f(a_i; N, P) \sum_{j=(1-P)\times N}^{N-1} \left( \frac{(N-1)!}{(N-1-j)! \times j!} \right) \left\{ j \times F(a_i; N, P) \right\}^{j-1} (1 - F(a_i; N, P))^{N-1-j} - \left[ (N-1-j)F(a_i; N, P) \right]^{j} (1 - F(a_i; N, P))^{N-2-j} \right\}. \tag{11}
$$

The solution to Equation (11) take the parameters $N$ and $P$, and calculates the equilibrium choice of output as a function of ability. Dividing the equilibrium score by the student’s ability returns the student’s choice of effort at equilibrium.

Now consider two different sizes of contest, $N$ and $N'$, each with the same proportion of winners, $P$. Evaluating Equation (11) at the different parameter values will reveal the effect that moving from a contest of size $N$ to size $N'$ has on the optimal effort of a student. Call this change the “treatment effect.”
4.1 Experimental Model

Solving for the optimal score and effort requires parameter values for \( N \) and \( P \). This subsection restricts attention to the experimental parameters, \( N = 10 \) and \( N = 100 \), and explores the effect that moving from one grading cohort size to another has on the effort exerted by students of different abilities at equilibrium. Let \( e^*_{i,t}(a_i) \) represent the equilibrium effort of a student with ability \( a_i \). Mathematically, I test \( \frac{\partial e^*_{i,t}(a_i)}{\partial N} \).

Each of the two grading cohort sizes are randomly assigned to quizzes, so call the two treatments the 10-Student Quiz and the 100-Student Quiz. The proportion of winners on each quiz is held constant at \( P = .7 \). That is, approximately 70 percent of the students receive high grades on a given quiz. Figure 1 maps a student’s ability to the probability they draw a grading cohort in which their ability places them in the top 70 percent as calculated by Equation (8).

Now, suppose that abilities are distributed uniformly over the unit interval and that marginal costs are constant at unity. That is, \( a_i \sim U[0,1] \), and \( C(e_{i,t}) = e_{i,t} \). Figure 2 plots the equilibrium score functions that solve Equation (11) for \( P = 0.7 \) and \( N = \{10, 100\} \) under these distributional and cost assumptions.

The intuition behind these equilibrium score functions can be gathered through a thought experiment. Consider the symmetric best response function in an environment where the proportion of winners remains constant at \( P = 0.7 \), but the number of students in a grading cohort approaches infinity. While a grading cohort of this size will still have some variability in the draw of students, the law of large numbers ensures that the realized distribution of students approaches a near-perfect reflection of the probability distribution from which they are drawn. Thus, common knowledge of the probability distribution is sufficient for a student’s belief about her relative position in her grading cohort to approach certainty.

In this infinitely large contest, a student whose ability is greater than the 30th percentile

\(^{12}\)With homogeneous marginal costs, any change in the marginal costs that does not result in a corner solution simply rescales the equilibrium scores.
Figure 1: Probability of Receiving Full Credit

in the probability distribution will best respond by choosing a score such that no student below the 30th percentile can match her score and receive non-negative expected surplus. That score, given the production and cost functions, is approximately $s_{i,t} = 0.3$.\footnote{This value is simply an itself means little, since it is determined by our choices of production and cost functions as well as the distribution of $a_i$.} Students below the 30th percentile best respond by exerting zero effort and achieving a score of zero. The equilibrium effort function in this setting, therefore, approaches a step function that starts at zero effort until $a_i \geq 0.3$, at which point the equilibrium effort jumps to generate a score of $s_{i,t} = 0.3$.\footnote{In the probability zero event that a student has exactly $a_i = 0.3$, he would need to be mixing such that students on either side of him would not wish to deviate from equilibrium} 

Keeping in mind the limiting case, consider the equilibrium scores in Figure 2. For the 100-Student Quiz, the scores will more closely reflect the infinitely large contest, while the
10-Student Quiz will be more affected by the randomness of its smaller grading cohort. This randomness manifests itself in the marginal costs and benefits of changing scores. For students who choose a score of $s_{i,t} = 0.3$ in the limit case, the marginal benefit of lowering a score is constant, since effort has a constant marginal cost, so foregone effort has a constant marginal benefit. A student’s marginal cost of lowering her score is paid through reductions in her probability of receiving a high grade. In the 10-Student Quiz, that probability changes more gradually, so, at the margin, reducing her score is less costly.

For the students with scores of zero in the infinitely sized contest, the cost of moving away from the equilibrium of the limiting case are paid in the form of positive effort. Increasing effort bears a constant marginal cost for both the 10- and 100-Student Quizzes, but holds a higher marginal benefit in the 10-Student Quiz for students with $a_i < 0.3$ because the
randomness increases the likelihood of states of the world in which even low ability students receive high grades.

Once equilibrium scores have been determined, equilibrium effort is deduced by dividing the equilibrium score by the ability of the student. My experiment pairs 10- and 100-Student Quizzes each week, and my analysis takes the difference in effort between the two quizzes as its dependent variable, so I will focus on the model’s predictions in this domain. Analyzing the difference within an individual controls for student-specific heterogeneity, and will provide a cleaner test of the treatment effect. To see the model’s predictions for how the difference in effort will evolve across the type-space consider Figure 3, which shows the equilibrium effort in the 100-Student Quiz minus the equilibrium effort in the 10-Student Quiz as a function of ability. I refer to this difference as the treatment effect.

![Figure 3: Difference in Equilibrium Effort](image-url)
There are two important features to note about Figure 3. First, the treatment effect crosses the horizontal axis exactly once. Meaning that the treatment effect is negative for all ability levels until near the critical cutoff of $a_i = 0.3$, after which the treatment effect is positive for all remaining ability levels. This predicts that the 10-Student Quiz will elicit greater effort from low ability students, and lower effort from high ability students. Second, the minimum and maximum treatment effects exist just on either side of the critical cutoff. This arises from the more abrupt changes in the equilibrium probabilities of receiving a high grade shown in Figure 1. Moldovanu and Sela (2006) prove these properties for a general class of cost functions. The characteristics arise out of a single-crossing property that they demonstrate must exist for symmetric equilibria in contests with different values of $N$, holding constant the proportion of winners.

5 Experimental Design

My experiment takes the experimental design used in Andreoni and Brownback (2014) and adapts it for a classroom context. My design simultaneously presents students with a pair of quizzes, the 10- and 100-Student Quizzes, and records behavior on each. The paired quizzes borrow from the paired auction design of Kagel and Levin (1993) and Andreoni, Che, and Kim (2007). In order to control for student-specific and week-specific effects, the difference in behavior between the two quizzes serves as the unit of analysis.

5.1 Recruitment and Participation

The experiment was conducted in the Winter quarter of 2014 in an intermediate economics classroom at The University of California, San Diego. Enrollment in the course started at 592 students, and ended at 563 after some students withdrew from the course. The experiment was announced both verbally and via web announcement at the beginning of the course. The announcement can be found in the appendix.
Students completed roughly 85 percent of the assigned quizzes in the course. The quizzes left incomplete were either a result of forgetfulness, apathy, or withdrawal from the course.

5.2 Quiz Design, Scoring, and Randomization

Over the course of the 10 week quarter there were 5 Quiz Weeks. At noon on Thursday of each Quiz Week, 2 different quizzes covering material from the previous week were posted on TED, the online course website used by UC San Diego. Both quizzes were due by 5pm the following day. Each quiz had a time limit of 30 minutes, and students could take the quizzes in any order they desired.

All students were assigned the same two quizzes each Quiz Week, call them Quiz A and Quiz B. The content of each quiz was designed to have as little overlap as possible with the other quiz to eliminate order effects in effort or scores. All students would also see both a 10- and 100-Student Quiz each week. The pair of quizzes was randomly assigned to the grading treatments. So, while all students saw the content from both Quiz A and B, approximately half had Quiz A graded as the 100-Student Quiz and the other half had Quiz A graded as the 10-Student Quiz, and oppositely with Quiz B. Before beginning the quiz, students only observed the grading treatment, and not the content of the quiz. No student was aware of the treatment received by any other student. Table 1 shows the balance across treatment.

<table>
<thead>
<tr>
<th>Quiz Version</th>
<th>Treatment</th>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
<th>Week 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10-Student</td>
<td>201</td>
<td>282</td>
<td>258</td>
<td>253</td>
<td>249</td>
</tr>
<tr>
<td></td>
<td>100-Student</td>
<td>262</td>
<td>282</td>
<td>259</td>
<td>253</td>
<td>243</td>
</tr>
<tr>
<td>B</td>
<td>10-Student</td>
<td>267</td>
<td>280</td>
<td>262</td>
<td>251</td>
<td>243</td>
</tr>
<tr>
<td></td>
<td>100-Student</td>
<td>200</td>
<td>282</td>
<td>259</td>
<td>256</td>
<td>252</td>
</tr>
</tbody>
</table>

Note: Asymmetries across treatments may arise out of chance, failed submissions, or withdrawals.

All quizzes were graded out of 3 possible points. Students receiving high grades were awarded 3 points, while students receiving low grades were awarded 1. Non-participants received 0 points.
The number of correctly answered questions determined the score for each student. The 10-Student Quiz featured 10 students in a grading cohort where the top 7 scores received high grades, and the 100-Student Quiz featured 100 students in a grading cohort where the top 70 scores received high grades. Students were randomly assigned to cohorts and all students in a cohort were assigned the same quiz under the same grading treatment. Students whose scores were tied at the cutoff were all awarded 3 points unless the tied students all failed to participate, in which case the students all received 0 points.

5.3 Effort

The TED site recorded the time at which every quiz was started and completed to the millisecond. The analysis will take the amount of time that a student spent taking the quiz to be the measure of effort that the student exerted on the quiz. This measure of effort will reveal which quiz the student believed held the greater returns to his effort.

Both quizzes were posted simultaneously, meaning that the time spent studying prior to either quiz was roughly constant between the two quizzes. This assertion is supported by the data, where 86 percent of students waited less than an hour between the two quizzes, with a median interval between quizzes of 32 minutes.

5.4 Ability

At the beginning of the course, students consented to the use of their grade point average in this study. This will serve as the measure of academic ability in the analysis. I elected not to use exam performance in the course because of the endogeneity between the allocation of effort to exams and quizzes.

Figure 4 shows the distribution of GPAs for students involved in the experiment. The value of the cumulative distribution function at a given GPA will represent the ability level of that student in the following predictions.

Importantly, the 30%-ile of the GPA histogram is located at 2.72. The median and mean
GPA are 3 and 2.99, respectively. 7 students have GPAs of 4.0, while only 1 student has the minimum GPA of 1.0.

5.5 Predictions From the Model

The model provides 3 primary predictions about the treatment effects. While the specific estimates of the model depend on too many assumptions to be particularly informative, the generic qualities will still provide clarity about the ways in which an expected-grade maximizing student would react to changes in their grading environment.
Hypothesis 1: The aggregate treatment effect if positive

My model predicts that aggregate effort will increase as the size of the grading cohorts increase. That is, the sign of the aggregate treatment effect should be positive. As cohort size increases, the law of large numbers decreases the uncertainty of the cohort distribution, increasing the probability that a student above the critical cutoff receives a high grade, causing those students to increase their effort. This increase more than offsets the decrease in effort by the students below the cutoff, for whom the probability of receiving a high grade decreases.

Hypothesis 2: The treatment effects cross the axis from below

Call the single-crossing point $a^*$. From Figure 3 it is clear that the treatment effect must be negative for ability levels less than $a^*$ and positive for ability levels greater than it. To see why, consider the limiting cases, $a_i = 0$ and $a_i = 1$. Regardless of the cohort size, students with $a_i = 0$ or $a_i = 1$ have probabilities of 0 and 1 of receiving a high grade, respectively. Therefore, students at either limit should be insensitive to treatment effects. Increasing ability from $a_i = 0$ draws the effort in the 10-Student Quiz above the effort of the 100-Student Quiz, generating a negative treatment effect up until $a^*$. Similarly, decreasing ability from $a_i = 1$ generates a positive treatment effect until $a^*$.

Let $T(a_i)$ represent the treatment effect. That is, $T(a_i) \equiv e_{i,t}(a_i; 100, 0.7) - e_{i,t}(a_i; 10, 0.7)$. Therefore,

$$T(a_i) < 0 \forall a_i \leq a^*$$
$$T(a_i) > 0 \forall a_i > a^*.$$

The 30th percentile of the GPA distribution corresponds to a GPA of 2.72. Since $a^*$ closely approximates the critical cutoff, the model predicts treatment effects to be negative.
for $GPA \leq 2.72$ and positive for $GPA > 2.72$.

**Hypothesis 3:** The local minimum is located just before $a^*$, and the local maximum is located just after $a^*$

From Equation (11), it is straightforward to derive the local extrema of $T(a_i)$ for given values of $N$ and $P$. The local minimum and maximum treatment effects appear just before and after the critical cutoff, respectively. The extrema identify the students for whom the relative returns to effort in one cohort size is maximally different from the corresponding returns in the other cohort size. Just before $a^*$, students see the greatest relative gains from the randomness of the 10-Student Quiz, while the opposite is true for students with ability just above $a^*$.

### 6 Results

I begin this section by describing my data, its basic statistics, and potential selection issues. Then, I specify the dependent variable I will use in the analysis and note its summary statistics. Next, I demonstrate heterogeneous treatment effects across students of different ability levels. Finally, I test the hypotheses of the model and catalogue regions over which the model does and does not hold predictive power.

#### 6.1 Data and Descriptive Statistics

In total, 579 students submitted 5,094 online quizzes in this experiment. Of those, 2,546 came from the 10-Student Quiz and 2,548 came from the 100-Student Quiz. The duration of each quiz was recorded. Table 2 reports the means and standard deviations of the unpaired quiz duration for both grading cohort sizes.

Quizzes were not perfectly equal in difficulty, and this affected the duration of each quiz. Since treatments were randomly assigned to quizzes, I will directly control for quiz-specific
Table 2: Descriptive statistics for duration of quiz

<table>
<thead>
<tr>
<th></th>
<th>100-Std. Quiz</th>
<th>10-Std. Quiz</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration (in minutes)</td>
<td>14.97</td>
<td>14.58</td>
<td>0.39</td>
<td>0.12</td>
</tr>
<tr>
<td>Standard Error</td>
<td>(0.18)</td>
<td>(0.18)</td>
<td>(0.26)</td>
<td></td>
</tr>
</tbody>
</table>

effects in all analysis.

6.2 Dependent Variable

The analysis will all be conducted using a student-specific difference between the two grading cohort sizes as the unit of analysis. This measure is appealing, because it reveals a student’s beliefs about which quiz will yield higher returns to his effort. Given the random assignment of treatments to quizzes, effort costs should be independent of the grading cohort size. Therefore, if a student spends more time on a given quiz, it is an indication that the student believes that his marginal product is higher on that quiz.

As the dependent variable, I will use a student-specific treatment effect. This effect will be the within-student difference between the effect of the 100-Student Quiz and the 10-Student Quiz on the amount of time the student spends on a quiz. This specification of the dependent variable will offer the best control for individual-specific noise in the data.

To generate student-specific treatment effects, I perform a first stage of analysis in which the duration of the quiz is regressed onto test difficulty controls and two fixed effects for each individual student, one for each size of grading cohort. This first stage is mathematically represented by,

\[ E_{i,t,N} = T_i \beta + \delta_{i,10} + \delta_{i,100} + u_{i,t} , \]

where \( \beta \) is the vector of coefficients capturing the difficulty of a given test, \( \delta_{i,10} \) represents student \( i \)'s fixed effect for 10-Student Quizzes, and \( \delta_{i,100} \) represents student \( i \)'s fixed effect for 100-Student Quizzes. Student \( i \)'s treatment effect is then defined as
$D_i = \delta_{i,100} - \delta_{i,10}$.

**Hypothesis 1: Aggregate treatment effect is positive**

Figure 5 captures the mean and the 95 percent confidence interval of the student-specific treatment effects.

On average, students spend approximately 27 additional seconds on the 100-Student Quiz than on the 10-Student Quiz, and this difference is statistically significant. When viewed relative to the means from Table 2, this difference is economically meaningful as well. A 27 second increase in time is more than a 3 percent increase over the mean. This confirms the prediction that the aggregate effort increases in the size of the grading cohort.
6.3 Endogenous Selection and Controls

One confound of these results is that students could choose the order in which they completed the quizzes. This was a limitation of the online environment, but the randomization process will provide a strong instrument to address this problem.\footnote{In order to force students to take quizzes in a specific order the second quiz must be hidden from view until the completion of the first quiz. Given the existing confusion over the quizzes, I felt that this strategy was infeasible.}

While students could choose to take the quizzes in any order they wanted, the order in which the quizzes were displayed was randomly assigned. So, while only 51.4 percent of the 100-Student Quizzes were presented first, students elected to take 55.3 percent of them before the corresponding 10-Student Quiz. Fortunately, there is a highly significant positive correlation between these two events, meaning that the randomly assigned order provides a relevant instrument that is mechanically designed to be valid as well.

Since my analysis is at a student-specific level, the confounding variable will be the percentage of 100-Student Quizzes taken first. To instrument for this choice, I use the percentage of 100-Student Quizzes displayed first to the students.

The first column of Table 3 captures the correlation and significance for the student-specific instrument.

<table>
<thead>
<tr>
<th>Table 3: Test of the Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pct of 100-Std. Quizzes taken first</td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td><strong>Pct of 100-Std. Quizzes displayed first</strong></td>
</tr>
<tr>
<td>GPA</td>
</tr>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>N</td>
</tr>
</tbody>
</table>

* p<0.10, ** p<0.05, *** p<0.01

The instrument, however, has a problematic correlation with the GPA of the student. Despite being randomly assigned before each Quiz Week, the ordering of the tests happened
to correlate to the GPA of the students. Column 2 of Table 3 captures this problematic correlation using a linear regression of the number of 100-Student Quizzes displayed first on the student’s GPA.

Column 3 offers some hope for the instrument. The results of a regression of the number of 100-Student Quizzes taken first on the number of 100-Student Quizzes displayed first and the student’s GPA show that the student’s GPA is not a significant predictor of the quiz order when controlling for the presentation order.

To demonstrate that none of these selection issues drive any results, all tables will feature results with and without direct controls for quiz order as well as the instrumented version of that control. I will note where the results substantively change because of the specification.

6.4 Heterogeneity

Figure 6 fits a local polynomial function mapping GPA to the duration of individual 10- and 100-Student Quizzes. One clear pattern is that the total amount of time allocated trends upwards with GPA. Additionally, the 100-Student Quiz duration crosses the 10-Student Quiz duration from below.

Absolute measures of the duration of quizzes are useful to display qualitative patterns of effort, but have significant problems with omitted variables that will affect both duration and GPA. In order to control for any individual level effects, histories of treated quizzes, and risk preference, my analysis will focus only on the difference between pairs of quizzes.

A first step in demonstrating strategic responding to changes in the environment is to demonstrate that students with different GPAs respond to changes in the size of the grading cohort in measurably different ways. To do this, I place students into three bins of equal length over the support of the GPA distribution. The bins are
Figure 6: Time spent on each quiz across student GPA’s
\[ Low_i = \mathbb{1}_{\{GPA \in [1, 2)\}} \]
\[ Mid_i = \mathbb{1}_{\{GPA \in [2, 3)\}} \]
\[ High_i = \mathbb{1}_{\{GPA \in [3, 4]\}} \].^{16}

Table 4 displays the results of regressing the student-specific treatment effects onto the bins with and without controls.

<table>
<thead>
<tr>
<th>Student-Specific Treatment Effects</th>
<th>2.103**</th>
<th>1.334*</th>
<th>2.103**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mid(_i)</td>
<td>(0.82)</td>
<td>(0.69)</td>
<td>(0.89)</td>
</tr>
<tr>
<td>High(_i)</td>
<td>1.542*</td>
<td>0.773</td>
<td>1.542*</td>
</tr>
<tr>
<td>Percent of 100-Std.</td>
<td>6.377**</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>Quizzes displayed first</td>
<td>(0.68)</td>
<td>(2.63)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-1.292*</td>
<td>-4.067***</td>
<td>-1.293</td>
</tr>
<tr>
<td>Instrumented</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>579</td>
<td>579</td>
<td>579</td>
</tr>
</tbody>
</table>

* p<0.10, ** p<0.05, *** p<0.01

From Table 4 it is clear that the treatment effect shows significant differences across ability levels regardless of the specification and controls implemented. The middle range of abilities has the strongest positive treatment effect—recall that this implies that they spend relatively greater amounts of time working on the 100-Student Quiz—while the low end of the distribution has the smallest treatment effect.

### 6.5 Relationship to Ability

This section details the predictability and policy-relevance of the patterns that arise in the treatment effects for students with different abilities. In addition, I test hypotheses 2 and
3 to uncover if these patterns coincide with the theoretical predictions. For this analysis, I divide the data into the following 4 bins,

\[
\begin{align*}
Bin 1_i &= \mathbb{1}_{\{GPA \in [1, 2)\}} \\
Bin 2_i &= \mathbb{1}_{\{GPA \in [2, 2.7)\}} \\
Bin 3_i &= \mathbb{1}_{\{GPA \in [2.7, 3.4)\}} \\
Bin 4_i &= \mathbb{1}_{\{GPA \in [3.4, 4]\}}.
\end{align*}
\]

These 4 bins provide enough variation to independently test Hypotheses 2 and 3 and make a semi-parametric characterization of the general patterns of student effort allocation.

Regressing the treatment effects onto the 4 bins generates the results displayed in Table 5.

<table>
<thead>
<tr>
<th>Table 5: Ability and the treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student-Specific Treatment Effects</td>
</tr>
<tr>
<td>( Bin_{1_i} )</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( Bin_{3_i} )</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( Bin_{4_i} )</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Percent of 100-Std. Quizzes First</td>
</tr>
<tr>
<td>Constant</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Instrumented</td>
</tr>
<tr>
<td>N</td>
</tr>
</tbody>
</table>

* p<0.10, ** p<0.05, *** p<0.01
Hypothesis 2: Treatment effects cross the axis from below

From Figure 4, the critical cutoff of this experiment corresponds to a GPA of approximately 2.72. Therefore, Hypothesis 2 predicts that the mean treatment effects for students below and above this cutoff should be negative and positive, respectively.

Since Bin2 is omitted, this requires a statistical test of the difference between the coefficient for Bin1 and the sum of the two coefficients for Bin3 and Bin4. This test fails to reject the null hypothesis that the two sums are the same ($F(1, 574) = 0.02$ $P = 0.90$). In fact, the mean difference between the 100- and 10-Student Quiz pairs below the critical cutoff, 47.1 seconds, is greater than the mean difference above the critical cutoff, 19.5 seconds.

Hypothesis 3: The local minimum is located just before $a^*$, and the local maximum is located just after $a^*$

Here I test the position of the local extrema relative to the critical cutoff. Specifically, I want to identify which bins are statistically distinguishable from the others in both the positive and negative directions.

A joint test of the difference between Bin2 and all other bins demonstrates that the treatment effect is highest in that region ($F(1, 574) = 5.64$ $P = 0.02$). Meanwhile, a joint test of the difference between Bin1 and all other bins shows that the treatment effect is lowest for students in that range, but this difference is not statistically significant ($F(1, 574) = 2.62$ $P = 0.11$). These results contradict Hypothesis 3, which predicts that the treatment effect should be minimized in Bin2 and maximized in Bin3.

6.6 Treatment Effect as a Continuous Function of Ability

Imposing continuity on the evolution of the treatment effect across abilities allows me to test the heterogeneity of the treatment effect across students of different abilities in a more structured way. Figure 7 plot local polynomial fits along with confidence intervals for the raw differences and the student-specific effects, respectively.
The figures confirm the patterns indicated by the previous tests. The peak of the predicted treatment effect occurs before the critical cutoff, where it is significant and positive, and trends downward towards 0 as a student’s ability increases. This fit also captures the negative treatment effect for low-ability students, but it is now statistically distinguishable from zero.

7 Discussion

Any discussion of effort exertion in classrooms first needs to address the social benefit of inducing greater effort on classroom quizzes. While it is not obvious that additional time spent on quizzes has any return in the form of performance or learning, what can certainly
be said is that additional time spent on quizzes is a measure of costly effort exertion. If students did not perceive it to increase performance, then any positive effort would be strictly dominated for any student with non-zero value of time. Accordingly, it serves as an appropriate proxy for the relative amount of effort a student would assign to each type of quiz in the absence of the paired quiz design. That is, if a student is willing to spend significantly more time on one quiz than the other when they are administered simultaneously, that same student would likely be willing to study longer or attend more class in preparation for the revealed preferred quiz in the absence of this experiment.

The connection between greater effort in studying or attendance and greater academic performance has been documented (Romer, 1993; Stinebrickner and Stinebrickner, 2008; De Fraja, Oliveira, and Zanchi, 2010; Arulampalam, Naylor, and Smith, 2012), implying that classroom motivation certainly generates positive returns. Understanding the strategic shifting of student effort resulting from seemingly small changes within a commonly applied grading scheme therefore serves to further the social goal of increasing academic output, and deserves attention as such.

Towards that end, my experiment uncovered systematic trends among students of different abilities to react to changes in the randomness of their grading environment that arise from changes in the number of students in their grading cohort. The advantage of my experimental design is that it can address the strategic effects of higher or lower enrollment in a course graded on the curve while holding constant all other characteristics of the classroom (such as teaching quality, student observability, or access to resources). With these clean controls, I can contribute to the discussion of optimal classroom size by demonstrating that changes in a students strategic setting causally affect the effort that a student exerts on that course.

\textsuperscript{17}Indeed, statistical tests fail to reject the null hypothesis that there is no relationship between the treatment and quiz scores of students.

\textsuperscript{18}Ultimately, however, this experiment was not designed to test the connection between the motivation to spend more time taking a given quiz and the motivation to study harder before quizzes, so I must leave that connection as well-founded speculation.
7.1 Patterns of Effort Allocation

In order to improve models of classroom effort allocation and make meaningful policy prescriptions based on observed patterns, three prominent stylized facts of the data need to be accounted for.

First, there was a positive average treatment effect. This implies that, on average, students increase their effort as the variance of their grading environment decreases. This result confirms the prediction from my model that when the variance of a grading environment decreases, the increases in effort from high ability students more than offset the decreases in effort from low ability students.

Second, there was a significant, negative treatment effect for students farthest below the critical cutoff. This indicates a preference among low-ability students to exert greater effort on the 10-Student Quiz than on the 100-Student Quiz.

Third, the peak of the treatment effect was located just before the critical cutoff. This was contrary to the model’s prediction that the returns to effort for those students should be higher in the 10-Student Quiz than in the 100-Student Quiz.

7.2 Misallocation of Effort

The significance the positive treatment effect for students just below the critical cutoff lies in the misallocation of effort between quizzes for those students. For students in the bottom 30th percentile, the higher variance environment offers them a greater opportunity to receive a favorable draw for their grading cohort. This randomness increases the returns to effort on the 10-Student Quizzes relative to the 100-Student quizzes for low ability students. Despite the greater returns to effort, students in the bottom 30th percentile did not take advantage of the randomness of the 10-Student Quiz. This misallocation increases the aggregate effort elicited by the 100-Student Quiz, at the cost of foregone potential grades for lower-ability students.

While the mechanisms for this misallocation of effort are not clear from the data, there
are several behavioral biases that could offer at least partial explanations. First, a failure of students to properly update about their environments or about their abilities could cause systematic tendencies towards one quiz over the other. Possibly arising from this first explanation is the second possibility. Students could possess a general overconfidence about their abilities. If students are strategically responding to their grading environment as if they possess their inflated estimate of their relative ability, then students just below the critical cutoff will favor the 100-Student Quiz, while students far below the critical cutoff will most favor the 10-Student Quiz. Both a failure to update and overconfidence are consistent with results from Eil and Rao (2011) and Mobius et al. (2011). Both of these studies find that agents fail to update about information over which they may have ego-utility.

A third explanation could come in the form of myopic loss aversion (Benartzi and Thaler, 1995). This loss aversion must be myopic, or else students would view increases in the probability of full credit on one quiz as perfectly substitutable for increases in the probability of full credit on another. Moreover, these reference points must be expected value based as in disappointment aversion (Bell, 1985; Loomes and Sugden, 1986; Gul,1991), because stochastic reference points in the style of Köszegi-Rabin (2006, 2007) suggest that students are risk neutral over changes in probability within the support of the original gamble. Thus, if a student has stochastic reference points, she would exert more effort on the quiz that led to the greatest increases in her probability of gains.

While I find these behavioral explanations compelling, my experiment was designed to uncover the allocation of effort by strategic students, not to tease apart each possible preference specification driving the allocations. As such, speculation about preferences will require further testing.

7.3 Policy Prescriptions

Based on the positive average treatment effects, a mechanism designer who has preferences only over aggregate effort would implement a grading environment with the lowest possible
variance in order to maximize the total effort exerted. This result could be accomplished through combining multiple classes into one grading unit after compensating for classroom level differences in teaching.

While decreasing the variance increases aggregate effort, it is not without certain costs. For the low ability students, it is clear that the higher variance environment induces more effort. The intuition for this result relates back to Figure 1, where it is clear that increases in the number of students in a grading cohort make it increasingly unlikely that a low-ability student receives a high grade. From a policy-perspective, this result suggests that low-ability students can become discouraged by relative grading when the cohort size becomes large enough. If motivating low-ability students is part of an educator’s objective function, then these students could benefit from being graded in smaller cohorts. This could be achieved by splitting large classes up into smaller sections and grading each relative to their own members.

The results also suggest that lower variance environments induce greater effort from students just below the critical cutoff. While the data do not explain the origin of this allocation failure, it still generates several policy prescriptions. If the designer hopes to motivate students just below the critical cutoff, then decreasing the variance of the grading environment appears to achieve that goal. It appears, however, that this comes at the cost of students farthest below the critical cutoff, so any mechanism designer will need to balance gains from the former group with losses from the latter.

These results also identify several failures of the model to capture all of the relevant phenomena. When taking into consideration behavioral biases, many different policy prescriptions arise. Failures to update about the environments or about their own ability could lead to misallocations of effort for students. Information about own ability or environmental factors could increase student utility in the event that they fail to make proper inference based on outcomes. For example, if students are over-confident, they could increase their own utility by best responding to their true ability level, implying that feedback may be welfare
increasing. At the same time, feedback may reduce aggregate effort, since the over-confidence of students below the critical cutoff added to the aggregate effort in the 100-Student Quiz. This may give rise to issues where the student is over-exerting effort, and the instructor wishes to enable this bias because it increases the amount of effort exerted in the class, causing even more complicated policy issues. The evidence of this phenomenon in my experiment is simply suggestive, and further experimentation is needed to confirm it. If students possess myopic reference points that cause them to allocate effort between different assignments in a way that is not grade-maximizing, then, again, information from instructors could encourage them to reallocate in ways that costlessly improve their grades.

8 Conclusion

In this paper, I uncovered heterogeneity in the way changes in class size affect students of different abilities when the class is graded on the curve. Understanding that students identify the classroom as a strategic environment should greatly benefit educators and administrators as they seek to design classroom environments and grading schemes to achieve their objectives with respect to student effort. While constraints will always exist with regards to the complexity of grading schemes and classroom environments, knowledge of the unintended consequences of these designs can greatly benefit students and teachers alike.

In order to ground the intuition for why class sizes may affect student effort choices, I first develop a theoretical model of the situation. This model takes into account the distribution of student abilities and the size of the class. Assuming students have grade-maximizing preferences, this model generates an understanding of the ways in which students would respond to shifts in the class size. I then conducted an experiment testing these qualitative predictions by measuring how actual students in a natural setting reacted to changes in their simulated class size. In a unique, paired quiz design, I was able to measure how much effort students spent on quizzes that only differed by the number of students in the grading cohort.
The results of the experiment confirmed that students do indeed identify the changes in the environment that result from shifting the number of students in a grading cohort. The mean effort exerted on the quiz with the larger grading cohort was significantly higher than the mean of the quiz with the smaller grading cohort, implying that lower-variance environments would promote more effort from students, on average.

In addition to generating statistically distinct mean effort, certain patterns clearly arose. The two most robust findings should provide guidance to mechanism designers as they seek to implement grading schemes that generate the desired distribution of effort from the students. The first finding was that low ability students exerted significantly greater effort on the quiz with the smaller grading cohort. The second finding was that students whose ability put them just beneath the critical cutoff for high grades exerted significantly greater effort on the quiz with the larger grading cohort. The first result suggests that smaller classes would promote effort from low-ability students as a result of their increases in variance, and the second result suggests that larger classes would promote more effort from students who believe themselves to be within reach of the critical cutoff.

The misallocation of effort represented by these two phenomena may not be completely atheoretical, as they are consistent with a number of possible behavioral biases. Non-Bayesian updating, overconfidence, and reference dependent utility could all generate these phenomena. Further experimentation is needed to confirm or reject these theories, however.

These results should make it clear that the relative grading mechanisms currently in place have many unintended consequences that are generated by potentially unavoidable changes in the class size. While designers may operate under constraints that may not allow them to fundamentally alter the grading environment, this information can still serve to justify giving greater attention or care to different demographics based on their predicted responses to the grading environment. Where such constraints do not exist, however, these results can provide the basis for exploration of optimal grading mechanisms that take into account both the proportion of winners and the size of the class in order to generate the desired effort from
students at each ability level. Towards this goal of optimal grading, this paper has identified regions over which effort exertion is particularly sensitive to changes in the environment and uncovered the predicted changes in effort over those regions.